5P-01

Deep Neural Network を用いた雑音抑圧及びブラインド音源分離手法の提案とその評価

橋本直矢† 野田邦昭† 中臺一博‡ 尾形哲也†

†早稲田大学基幹理工学部表現工学科尾形研究室 ‡HRI-JP

1. はじめに

実環境下でのロボットと人間との円滑なコミュニケーションを考える時，雑音に頑健な音声認識の実現が重要になるが，その有効なアプローチとしてブラインド信号音源分離（BSS）が挙げられる．本研究では，BSS における分離フィルタ・雑音抑圧フィルタのモデルとして，高次の非線形写像を表現できる Deep Neural Network(DNN)を用い，従来の線形写像を用いた音源分離手法よりも高い性能を実現できることを示す．具体的にはマイクロホンアレイにより収録した混合音声信号の多チャンネルメルフィルタバンク特徴を入力，原信号である音源の特徴量を出力として DNN を教師あり学習し分離フィルタをモデル化する．このモデルの評価のため，孤立単語と大語彙連続発話の音声認識タスクを行い，従来手法や DNN を単純な雑音抑圧フィルタとして用いた場合等との性能比較を行う．

2. 音源分離の従来方法とその課題

2.1 独立成分分析による音源分離

収音された複数の未知の混合音声信号から，それぞれの音声を分離することをブラインド信号音源分離(BSS)という．その代表的な解法に，独立成分分析（ICA）がある．ICA では N 個の観測 $\mathbf{x}(t)$ が統計的に独立な M 個の原信号 $\mathbf{s}(t)$ の線形重ね合わせという仮定で，T 個の観測信号 $\mathbf{x}(t)$ が得られた場合に，$\mathbf{x}(t)=\mathbf{H}\mathbf{s}(t)$ のように混合行列 $\mathbf{H}$ を定義する．また信号分離は観測データ $\mathbf{x}(t)$ を分離信号 $\mathbf{y}(t)$ に変換するように分離式 $\mathbf{y}(t)=\mathbf{W}\mathbf{x}(t)$ によって実現される．

2.2 動的環境を考慮した音源分離と音声強調

中臺らは，ロボット聴覚など動的変化への追従性が求められる環境下で，Geometric High-order Decorrelation-based Source Separation with Adaptive Stepsize control (GHDSS-AS，以下 GHDSS と表記)という音源分離手法を提案している[1]．GHDSS とはマイクロホンアレイを利用し音源信号間の無相関化と音源方向への指向性の形成を行う BSS とビームフォーマの混合手法である．さらに，動的環境を考慮した音声強調を行うための雑音抑圧法として Histogram-based Recursive Level Estimation （HRLE）を提案している．

2.3 課題

基本的に GHDSS は線形分離手法なので，非線形混合は扱えないという点に課題がある．一方で，非線形混合を含む問題の解決のため，観測信号の生成モデルを仮定した上で，最尤推定によって混合行列や独立信号を推定するアプローチも提案されているが[2]，モデルの表現能力やスケーラビリティに限界があった．

3. DNN による音源分離と雑音抑圧

3.1 Deep Neural Network(DNN)

近年機械学習の分野で注目されている学習アルゴリズムに Deep Neural Network （DNN）がある．DNN は一般に神経回路モデルの階層を深くしたモデルのことで，高次の非線形写像を表現可能な点が特徴である．DNN は，大量の学習データから汎化性能の高い特徴量を自己組織化できることが特長で，音声認識タスクにおいて従来手法による特徴量を超える認識性能を発揮できることが報告されている[3]．また，雑音除去のモデルとしても高い性能を発揮できることが知られている[4]．そこで本研究では，DNN を音源分離フィルタ及び雑音抑圧フィルタのモデルとして応用し，従来手法を上回る音源分離及び雑音抑圧性能を実現することを目指した．

3.2 提案モデル

本研究では，8ch の混合音声信号のメルフィルタバンク特徴から 1ch のクリーンな音響特徴を出力するよう DNN を教師あり学習して分離フィルタのモデルとした．DNN は図 1 に示すように，入力層が隠れ層を通し徐々に次元圧縮され出力される構造とした．

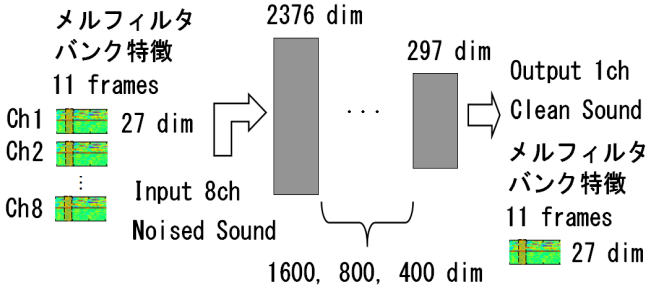

図 1:提案モデル

Blind Sound Source Separation and Denoising with Deep Neural Network: Naoya Hashimoto†, Kuniaki Noda†, Kazuhiro Nakadai‡ ,Tetsuya Ogata†

†Waseda University ‡HRI-JP

4. 実験設定

4.1 データセット

提案モデルの評価のため，孤立単語及び大語彙連続発話認識を行った．孤立単語は ATR 音素バランス単語 216 語を，大語彙連続発話は JNAS 新聞記事読み上げコーパスを用いた．さらに発話と混合される雑音は 2 つの条件を設定し，ロボットのファン雑音及び音楽雑音を用いた．

4.2 録音条件

残響時間約 0.2 秒，4x7m の室内で研究用ロボット（Hearbo）に付属するマイクロホンアレイを用いて収録を行った．音声発話はマイク正面から 30 度の位置，音楽雑音は 330 度の位置に設定し，距離 1m の距離からのインパルス応答を収録して雑音を合成した．なお，雑音レベルは 0，6，12dB の 3 種類を用意し，ファン雑音は方向性のない雑音とし，音楽雑音と同レベルとなるように合成した．

図 2: Hearbo

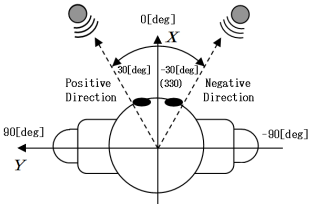

図 3:座標系

4.3 DNN の学習条件・構造

本実験では，DNN の入出力及び音声認識のデータ形式は 27 次元のメルフィルタバンク特徴を用い，入出力は 11 フレーム分の特徴量を 1 単位と扱ったため，入力は 2376 次元，出力は 297 次元となった．隠れ層の次元は 1600，800，400 の 3 層で，各層の活性関数はロジスティック関数を用いた．DNN の学習データは，JNAS コーパス(計 64938 発話分）を用い，特徴量に変換後その内 40 万サンプルをランダムに抽出し学習した．

4.4 比較条件

認識性能の比較実験として 6 種類の手法で音響特徴を準備した．それぞれ記すと，雑音の載ったそのままの音響データ（Noise），GHDSS で音源分離を行ったもの（GHDSS），GHDSS に加え，HRLE で雑音抑圧を行ったもの（HRLE），提案手法により音源分離を行ったもの（DNN），GHDSS，HRLE の処理結果から DNN でクリーンな音響特徴を予測したものとなっている（それぞれ，DNN（GHDSS），DNN（HRLE））．なお，音源のクリーンな音響データの認識率は孤立単語，JNAS でそれぞれ 99.00%，88.79%となった．

4.4 結果・考察

実験で得られた音響特徴を音声認識エンジン Julius で認識を行った結果を表 1 に示す．

今回の結果を全体的に見ると，音楽雑音を含む時の方がファン雑音の時よりも DNN がより有効という傾向が見られる．これは前述のように音楽雑音には方向性があるが，ファン雑音には無いため 8ch のマイクロホンアレイの使用による差異が出づらいことが原因と推測される．これより，ファン雑音では雑音レベルの高い領域では DNN の使用に優位性があるが，雑音レベルが低い領域では，GHDSS や HRLE 処理などの音源分離手法と性能に大きな差がないものと考えられる．

表 1: 認識率

孤立単語認識率(Fan Noise)

	Noise	GHDSS	HRLE	DNN	DNN (GHDSS)	DNN (HRLE)
0dB	40.00	83.61	85.28	86.44	86.99	★90.14
6dB	81.67	94.72	★95.14	94.17	93.33	95.05
12dB	95.51	★97.64	96.99	96.44	94.72	96.30

孤立単語認識率(Music Noise)

	Noise	GHDSS	HRLE	DNN	DNN (GHDSS)	DNN (HRLE)
0dB	25.32	64.31	74.44	★86.99	79.72	80.69
6dB	58.33	84.44	91.16	★94.12	92.18	92.50
12dB	79.40	89.54	94.12	★96.53	95.51	95.74

JNAS認識率 (Fan Noise)

	Noise	GHDSS	HRLE	DNN	DNN (GHDSS)	DNN (HRLE)
0dB	6.87	40.56	46.81	50.06	★51.79	50.02
6dB	34.60	69.44	★73.21	72.70	71.18	70.35
12dB	63.07	79.75	★81.29	79.68	77.85	77.85

JNAS認識率 (Music Noise)

	Noise	GHDSS	HRLE	DNN	DNN (GHDSS)	DNN (HRLE)
0dB	6.15	29.54	30.27	★51.63	34.53	34.45
6dB	21.66	58.17	61.33	★72.89	63.29	60.78
12dB	50.34	73.92	76.21	★79.55	74.85	74.43

5. 結論

本稿では，DNN を用いたブラインド音源分離手法を提案した．従来方法との認識率比較により，方向性のある音楽雑音が重畳されている場合，特に SN 比が小さい時に最も DNN による音源分離手法の優位点が発揮されることがわかった．

謝辞：本研究は，さきがけ領域研究「情報環境と人」及び科研費新学術領域研究「構成論的発達科学」(24119003）の助成を受けた．

参考文献

[1] K. Nakadai et al., “Sound source separation and automatic speech recognition for moving sources,” IEEE/RSJ IROS2010, pp.976–981, 2010.

[2] 前田新一, 石井信, “非線形雑音付き独立成分分析” 電子情報通信学会技術研究報告. NC, ニューロコンピューティング, pp.41–46, 2004.

[3]Geoffrey Hinton et al., “Deep Neural Networks for Acoustic Modeling in Speech Recognition” IEEE Signal Processing Magazine, pp.82-97, 2012.

[4]Xue Yang et al., “Speech Feature Denoising and Dereverberation via Deep Autoencoders for Noisy Reverberant Speech Recogniton” IEEE,ICASSP2014, pp.1778-1782, 2014.